Submitted exclusively to the *Journal of Mathematics and Music*
Last compiled on August 9, 2016

# Mapping between dynamic markings and performed loudness: A machine learning approach

Katerina Kosta[a*], Rafael Ramirez[b], Oscar F. Bandtlow[c] and Elaine Chew[a]

[a] *Centre for Digital Music, Queen Mary University of London, London, UK;*
[b] *Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain;*
[c] *School of Mathematical Sciences, Queen Mary University of London, London, UK.*

()

Loudness variation is one of the foremost tools for expressivity in music performance. Loudness is frequently notated as dynamic markings such as ***p*** (*piano*, meaning soft) or ***f*** (*forte*, meaning loud). While dynamic markings in music scores are important indicators of how music pieces should be interpreted, their meaning is less straightforward than it may seem, and depends highly on the context in which they appear. In this article, we investigate the relationship between dynamic markings in the score and performed loudness by applying machine-learning techniques—decision trees, support vector machines, artificial neural networks, and a k-nearest neighbor method—to the prediction of loudness levels corresponding to dynamic markings, and to the classification of dynamic markings given loudness values. The methods are applied to forty-four recordings of performances of Chopin's Mazurkas each by eight pianists. The results show that loudness values and markings can be predicted relatively well when trained across recordings of the same piece, but fail dismally when trained across the pianist's recordings of other pieces, demonstrating that score features may trump individual style when modeling loudness choices. Evidence suggests that all the features chosen for the task are relevant, and analysis of the results reveal the forms (such as the return of the theme) and structures (such as dynamic marking repetitions) that influence predictability of loudness and markings. Modeling of loudness trends in expressive performance appears to be a delicate matter, and sometimes loudness expression can be a matter of the performer's idiosyncracy.

**Keywords:** dynamic markings; loudness level representation; machine learning; loudness prediction; marking classification

## 1. Introduction

Information based on loudness changes abstracted from music signals can serve as an important source of data for the creation of meaningful and salient high-level features of performed music, which is almost all of the music that we hear. Our focus is on concepts related to dynamic levels in musical expressivity which are represented in the score by markings such as ***p*** (*piano*, meaning soft) and ***f*** (*forte*, meaning loud). These symbols are interpreted in performance and communicated through the varying of loudness levels. Understanding the relationship between dynamic markings and loudness levels is critical to applications in music cognition, musicology (performance analysis), and music informatics (e.g. transcription).

Our approach is a computational one. Existing work on computational modeling of

*Corresponding author. Email: katerina.kosta@qmul.ac.uk.com

loudness in performed music is limited and has largely focused on expressive performance rendering. Here, loudness is considered as part of the performances characteristics that have a large impact on the quality of the rendered pieces. One example is the YQX probabilistic performance rendering system based on Bayesian network theory described by **?**. In this study, loudness is, among others, one of the expressive performance rendering targets. A similar approach for defining loudness as a parameter of expression was taken by **?** in their statistical modeling of polyphonic piano renditions.

Other researchers have studied dynamics as an important tool for shaping performance using computational techniques. Widmer and Goebl (2004) reviewed various uses of Langner's tempo-loudness space in the representation and shaping of expressive performance parameters. Sapp (2008) used scapeplots to represent loudness variation at different hierarchical levels and to distinguish between performance styles.

Little work exists on score-based loudness feature representation, an exception being the study of Kosta, Bandtlow, and Chew (2014) on the meanings of dynamic markings in performances of five Chopin Mazurkas. A related study is that of Grachten and Krebs (2014) where a machine-learning approach is applied to score-based prediction of note intensities in performed music. From the perspective of transcribing the loudness changes to dynamic markings, the MUDELD algorithm proposed by **?** uses linguistic description techniques to categorize the dynamic label of separate musical phrases into three levels labeled as *piano* (***p***), *mezzo* (***m***), and *forte* (***f***).

In this paper we present a number of machine-learning approaches with a two-fold purpose: predicting loudness levels given dynamic score markings and classifying performed loudness into dynamic markings. This will be done by taking into account features that emphasize the fact that dynamic markings have to be understood in relative terms. The approach to the analysis of dynamic markings is inspired by the description of dynamics by Khoo (2007) as operating on "primary dynamic shading" (absolute) and "inner shadings" (relative) levels. A consequence is that the absolute loudness of dynamic markings may be superseded by their local context so that a ***p*** dynamic might objectively be louder than an ***f*** at another part of the piece, as shown by Kosta, Bandtlow, and Chew (2014).

Fabian, Timmers, and Schubert (2014) posit that while a significant part of the performer's aim is to communicate the composer's intentions, nevertheless, performers bring their personal, cultural, and historical viewpoints to the fore when subjectively understanding expression. In an early study on dynamics performance styles, **?** alludes to the difficulty of grouping patterns of dynamic changes across recordings of the same musical excerpt, highlighting that only a weak relationship exists between performers' sociocultural variables with the dynamics profile in their recordings. In order to systematically investigate variability in the performer's understanding of the composer's intentions, we analyze eight different pianists' recordings of forty-four distinct Mazurkas by Frédéric Chopin.

The remainder of the paper is organized as follows: In Section 2 we describe the data set used for this study, the features extracted, the learning task and the algorithms employed. Section 3 presents and discusses the results obtained when predicting loudness levels at dynamic markings; and, Section 4 does the same for results obtained when classifying loudness levels into dynamic markings. Finally, Section 5 summarizes the conclusions with some general discussions.

## 2. Material and methods

This section describes the music material and computational methods that form the basis of the studies of this paper. Section 2.1 describes the dataset, the multiple alignment strategy used to synchronize audio and score features, and the features extracted. Section 2.2 presents the machine-learning algorithms employed in the experiments.

### 2.1. *Data preparation*

For the purpose of this study, we examine recordings of eight pianists' performances of forty-four performances of Mazurkas by Frédéric Chopin. The Mazurkas and the number of dynamic markings each are detailed in Table 1, and the eight pianists together with the recording's year and index are identified in Table 2. The audio data was obtained from the CHARM Project's Mazurka database[1].

| Mazurka index | M06-1 | M06-2 | M06-3 | M07-1 | M07-2 | M07-3 | M17-1 | M17-2 | M17-3 | M17-4 | M24-1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # markings | 18 | 13 | 22 | 13 | 13 | 18 | 7 | 6 | 9 | 7 | 4 |
| Mazurka index | M24-2 | M24-3 | M24-4 | M30-1 | M30-2 | M30-3 | M30-4 | M33-1 | M33-2 | M33-3 | M33-4 |
| # markings | 12 | 7 | 33 | 8 | 14 | 25 | 18 | 5 | 16 | 4 | 12 |
| Mazurka index | M41-1 | M41-2 | M41-3 | M41-4 | M50-1 | M50-2 | M50-3 | M56-1 | M56-2 | M56-3 | M59-1 |
| # markings | 12 | 5 | 6 | 7 | 15 | 14 | 17 | 14 | 7 | 16 | 8 |
| Mazurka index | M59-2 | M59-3 | M63-1 | M63-3 | M67-1 | M67-2 | M67-3 | M67-4 | M68-1 | M68-2 | M68-3 |
| # markings | 8 | 11 | 9 | 4 | 18 | 10 | 13 | 11 | 12 | 21 | 8 |

Table 1. Chopin Mazurkas used in this study and the number of dynamic markings that appear in each one. Mazurkas are indexed as "M<*opus*>-<*number*>."

| Pianist | Chiu | Smith | Ashkenazy | Fliere | Shebanova | Kushner | Barbosa | Czerny |
|---|---|---|---|---|---|---|---|---|
| Year | 1999 | 1975 | 1981 | 1977 | 2002 | 1990 | 1983 | 1989 |
| ID | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 |

Table 2. Pianist's name, year of the recording, and pianist ID.

The loudness time series is extracted from each recording using the *ma_sone* function in Elias Pampalk's Music Analysis toolbox[2]. The loudness time series is expressed in sones, and smoothed by local regression using a weighted linear least squares and a 2nd degree polynomial model (the "loess" method of MATLAB's *smooth* function[3]). The final values are normalized by dividing each value with the largest one per recording; in this way we are able to compare different recording environments.

#### *2.1.1. Audio recording alignment*

To speed the labour-intensive process of annotating beat positions for each recording, only one recording (which we refer to as the "reference recording") for a given Mazurka was annotated manually, and the beat positions transferred automatically to the remaining recordings using a multiple performance alignment heuristic to be described below. The multiple performance alignment heuristic employs the pairwise alignment algorithm by Ewert, Müller, and Grosche (2009), which is based on Dynamic Time Warping (DTW)

[1] www.mazurka.org.uk, accessed 20 February 2016.

[2] www.pampalk.at/ma/documentation.html, accessed 20 February 2016.

[3] http://uk.mathworks.com/help/curvefit/smooth.html?refresh=true, accessed 20 February 2016.

applied to chroma features. This pairwise alignment technique extends previous synchronization methods by incorporating features that indicate onset positions for each chroma. Ewert et al. report a significant increase in alignment accuracy resulting from the use of these chroma-onset features and the average onset error for piano recordings is 44 milliseconds.

The pairwise alignment algorithm creates a match between two audio files, say $i$ and $j$, using dynamic time warping. The matched result is presented in the form of two column vectors $\mathbf{p}_i$ and $\mathbf{q}_j$, each with $m$ entries where $m$ depends on the two recordings chosen, $i$ and $j$. Each vector presents a nonlinear warping of the chroma features for the corresponding audio file, and represents the timing difference between the two recordings. A pair of entries from the two vectors gives the indices of the matching time frames of the two audio files. We compute the Euclidean distance between each pair of the dynamic time warped audio files as follows:

$$d_{i,j} = \sqrt{\sum_{k=1}^{m}(q_{j,k} - p_{i,k})^2}, \qquad \forall i \neq j, \tag{1}$$

where $m \in \mathbb{N}$ is the size of the vectors. In this way, each audio has a profile corresponding to its alignment to all other audio recordings which is $\mathbf{d}_i = [d_{i,j}]$. The average value of all the alignment accuracies for the $i^{th}$ recording in relation to the remaining ones is $\overline{\mathbf{d}}_i$.

The goal of the multiple performance alignment heuristic is to optimize the choice of a *reference audio* with which we can obtain better alignment accuracies than with another reference audio file. We consider the best reference file to be one that minimizes the average distance to other audio files and without extreme differences from more than two other audio recordings as measured by the norm distance. Mathematically, the problem of finding the reference audio can be expressed as one of solving the following problem:

$$\begin{aligned} & \min_i \overline{\mathbf{d}}_i \\ \text{s.t.} \quad & \#\left\{j : |d_{i,j}| > q_3(\mathbf{d}_i) + 1.5[q_3(\mathbf{d}_i) - q_1(\mathbf{d}_i)]\right\} \leq 2, \end{aligned}$$

where $q_\ell(\mathbf{d}_i)$ is the $\ell$-th quantile of $\mathbf{d}_i$, and the left hand side of the inequality uses an interquartile-based representation of an outlier. The "reference recording" is then given by $\arg\min_i \overline{\mathbf{d}}_i$.

We then detect manually the beat positions only for the reference audio recording and we obtain the beat positions of the remaining recordings by using the alignment method which is mentioned above. In order to evaluate the method, we have compared these derived beat positions with our manually annotated ones for forty-four recordings of the Mazurka Op. 6 No. 2 and the average error was 37 milliseconds.

### *2.1.2. Observed dynamic features*

The dynamic markings are taken from the Paderewski, Bronarski, and Turczynski (2011) edition of the Chopin Mazurkas. For the dataset described in Table 1, ***pp*** occurs 62 times, ***p*** 234, ***mf*** 21, ***f*** 170, and ***ff*** 43 times, giving a total of 530 dynamic markings. The loudness value corresponding to each marking is the average of that found at the beat of the marking and the two subsequent beats. More formally, if $\{y_n\} \in \mathbb{R}$ is the sequence of loudness values in sones for each score beat indexed $n \in \mathbb{N}$ in one piece, then the loudness value associated with the marking at beat $b$ is $\ell_b = \frac{1}{3}(y_b + y_{b+1} + y_{b+2})$.

About the choice of three beats, sometimes the actual change in response to a new dynamic marking does not take place immediately, and can only be observed in the subsequent beat or two. It is clear from the data that loudness varies considerably between one dynamic marking and the next. Thus, we additionally aim to have the smallest window possible to capture dynamic changes in response to a marking. Consequently, we have chosen a three-beat window as the window for study, which for Mazurkas corresponds to a bar of music.

For each dynamic marking in the data set, we have extracted the following associated features:

(1) label of the current dynamic marking (M);
(2) label of the previous dynamic marking (PR_M);
(3) label of the next dynamic marking (N_M);
(4) distance from the previous dynamic marking (Dist_PR);
(5) distance to the next dynamic marking (Dist_N);
(6) nearest non-dynamic marking annotation between the previous and current dynamic marking, e.g. ***crescendo*** (Annot_PR);
(7) nearest non-dynamic marking annotation between current and next dynamic marking, e.g. ***crescendo*** (Annot_N); and,
(8) any qualifying annotation appearing simultaneously with the current dynamic marking, e.g. ***dolcissimo*** (Annot_M).

In addition to the feature set described above, we also have an associated loudness value, L, which is the $\ell_b$ value on the beat of the dynamic marking.

We consider at most one value each for the features "Annot_ PR," "Annot_N," and "Annot_M." If two different annotations occur on the same beat, we choose the one related to dynamic changes, an exception being the case where the annotations ***sf*** and ***a tempo*** appear simultaneously. In this case, to be consistent with the time range of other non-dynamic marking annotations, we choose ***a tempo*** over ***sf***, as it applies to more than one score beat. In the case where there was an annotation related to change in dynamics and a qualifying term such as ***poco*** preceding it, we use the annotation without the qualifier, to limit the number of dynamic terms.

### 2.2. *Learning task and algorithms*

In this article we explore different machine learning techniques to induce a model for predicting the loudness level at particular points in the performance. Concretely, our objective is to induce a regression model M of the following form:

$$M(FeatureSet) \rightarrow Loudness$$

Where $M$ is a function which takes as input the set of features (*FeatureSet*) described in the previous section, and *Loudness* is the predicted loudness value. In order to train $M$ we have explored the following machine learning algorithms (as implemented in Hall et al. (2009)):

**Decision Trees (DT).** Decision trees Quinlan (1986) use a tree structure to represent possible branching on selected attributes so as to predict an outcome given some observed features. The decision tree algorithm recursively constructs a tree by selecting at each node the most relevant attribute. The selection of the most relevant attribute, at each node of the tree is based on the *information gain* associated with each attribute and

the instances at each node of the tree. For a collection of loudness values, suppose there are b instances of class B and c instances of class C. An arbitrary object will belong to class B with probability $b/(b+c)$ and to class C with probability $c/(b+c)$. The expected information needed to generate the classification for the instance is given by

$$I(b,c) = -\left(\frac{b}{b+c}\log_2\frac{b}{b+c} + \frac{c}{b+c}\log_2\frac{c}{b+c}\right). \tag{2}$$

Suppose attribute A can take on values $\{a_1, a_2, ..., a_v\}$ and is used for the root of the decision tree, partitioning the original dataset into $v$ subsets, with the $i$-th subset containing $b_i$ objects of class B and $c_i$ of class C. The expected information required for the $i$-th subtree is $I(b_i, c_i)$, and the expected information required for the tree with A as root is the weighted average

$$E(A) = \sum_{i=1}^{v} \frac{b_i + c_i}{b+c} I(b_i, c_i), \tag{3}$$

where the weight for the $i$-th subtree is the proportion of the objects in the $i$-th subtree. The information gained by branching on A is therefore gain$(A) = I(b,c) -$ E$(A)$, and the attribute chosen on which to branch at the next node is thus $\arg\max_A gain(A)$. Decision trees can also be used for predicting numeric quantities. In this case, the leaf nodes of the tree contain a numeric value that is the average of all the training set values. Decision trees with averaged numeric values at the leaves are called *regression trees*.

**Support Vector Machines (SVM).** Support vector machines by Cristianini and Shawe-Taylor (2000) aim to find the hyperplane that maximizes the distance from the nearest members of each class, called support vectors. Cristianini et al. use a nonlinear function $\phi(\cdot)$ to map attributes to a sufficiently high dimension, so that the surface separating the data points into categories becomes a linear hyperplane. This allows the model to predict nonlinear models using linear methods. The data can be separated by a hyperplane and the support vectors are the critical boundary instances from each class.

The process of finding a maximum margin hyperplane only applies to classification. However, support vector machine algorithms have been developed also for regression problems, i.e. numeric prediction, that share many of the properties encountered in the classification case. SVM for regression problems also produce a model that can usually be expressed in terms of a few support vectors and can be applied to nonlinear problems using kernel functions.

**Artificial Neural Networks (ANN).** Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. They are based on the idea of the perceptron which includes an input, a hidden and an output layer of data points connected with nodes having numeric weights. The nodes of the input layer are passive, meaning they do not modify the data. The two aspects of the problem is to learn the structure of the network, and to learn the connection weights. A multilayer perceptron (MLP)–or ANN– has a linear activation function in all neurons, that is, a linear function that maps the weighted inputs to the output of each neuron. More formally, we assume that in one network there are $d$ inputs, $M$ hidden units and $c$ output units. As it is described by **?**, the output of the $j$th hidden unit is

obtained by first forming a weighted linear combination of the $d$ input values, to give

$$a_j = \sum_{i=1}^{d} w_{ji}^{(1)} x_i, \tag{4}$$

where $w_{ji}^{(1)}$ denotes a weight in the first layer, going from input $i$ to hidden unit $j$. The activation of hidden unit $j$ is then obtained by transforming the linear sum above using an activation function $g(\cdot)$ to give

$$z_j = g(a_j). \tag{5}$$

The outputs of the network are obtained by transforming the activations of the hidden units using a second layer of processing elements. For each output unit $k$, we construct a linear combination of the outputs of the hidden units of the form

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j. \tag{6}$$

The activation of the $k$th output unit is then obtained by transforming this linear combination using a non-linear activation function $\tilde{g}(\cdot)$, to give

$$y_k = \tilde{g}(a_k). \tag{7}$$

We next consider how such a network can learn a suitable mapping from a given dataset. Learning is based on the definition of a suitable error function, which is minimized with respect to the weights and biases in the network. If we define a network function, such as the sum-of-squares error which is a differentiable function of the network outputs, then this error is itself a differentiable function of the weights. We can therefore evaluate the derivatives of the error with respect to the weights, and these derivatives can then be used to find weight values which minimize the error function.

In this paper, in order to determine the weights, that is, to tune the neural network parameters to best fit the training data, we apply the gradient descent back propagation algorithm of Chauvin and Rumelhart (1995). The back propagation algorithm is used for evaluating the derivatives of the error function and learns the weights for a multi-layer perceptron, given a network with a fixed set of units and interconnections. The idea behind this algorithm is that the output corresponds to a propagation of errors backwards through the network. We empirically set the momentum applied to the weights during updating to 0.2 and the learning rate, that is the amount of the weights that are updated, to 0.3. We use a fully-connected multi-layer neural network with one hidden layer meaning that we have one input and one output neuron for each attribute.

**k-Nearest Neighborhood (k-NN)**. k-NN is a type of instance-based learning which instead of performing explicit generalization, compares the new data instances with instances in the traning set previously stored in memory. In k-NN generalization beyond the training data is delayed until a query is made to the system. The algorithm's main parameter is $k$, the number of considered closest training vectors in the feature space. Given a query, the output's value is the average of the values of its $k$ nearest neighbors. More formally, as it is described by **?**, the k-nearest neighbor fit for the prediction $\hat{Y}$ of

the output $Y$ is

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i, \tag{8}$$

where $N_k(x)$ is the neighborhood of the input instance x defined by the closest points $x_i$ in the training sample.

In this paper we report the results of the $k$-NN algorithm with $k = 1$ which finds the training instance closest in Euclidean distance to the given test instance. If several instances are qualified as the closest, the first one found is used. We tested the values of $k$ equal to 1,2, and 3 and there was not a significant difference on the regression results.

## 3. Performed loudness-level modeling

This section assesses the fit of the machine-learned models as they predict loudness values in a pianist's recording of a Mazurka, first given other recordings of that Mazurka, and second given other recordings of that pianist. Two experiments have been conducted for this purpose. In the first one, each prediction model has been trained for each Mazurka separately. Then, each model has been evaluated by performing a 8-fold training-test validation in which instances of one pianist of the training set are held out in turn as test data while the instances of the remaining seven pianists are used as training data. In the second one, each prediction model has been trained for each pianist separately. Then each model has been evaluated by performing a 44-fold training-testing validation in which instances of one Mazurka of the training set are held out in turn as test data while the instances of the remaining forty-three Mazurkas are used as training data.

Section 3.1 presents the machine-learning algorithms' results of the first experiment, when predicting loudness given other pianists' recordings of the target piece; Section 3.2 presents the machine-learning algorithms' results of the second experiment, when predicting loudness given the target pianist's other recordings; Section 3.3 considers the extremes in prediction results for the second experiment; Section 3.4 considers the degree of similarity in approach to loudness between pianists; and, Section 3.5 considers the relevance of the features selected.

### 3.1. *Predicting loudness given other pianists' recordings of target piece*

In the first experiment, we use the machine-learning methods described above to predict the loudness values at the dynamic markings of one *Mazurka*, given the loudness levels at the markings of the same *Mazurka* recorded by other pianists. We test the machine-learning models by assessing their predictions at points where the composer (or editor) has placed a dynamic marking because this is our closest source of ground truth. The evaluations focus on how well a pianist's loudness choices can be predicted given those of other seven pianists for the same Mazurka.

As a measure of accuracy, we compute the Pearson correlation coefficient between predicted ($X$) and actual ($Y$) loudness values using the formula

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1} n(x_i - \overline{X})^2}\sqrt{\sum_{i=1} n(y_i - \overline{Y})^2}}, \tag{9}$$

where $x_i \in X$, $y_i \in Y$, and the size of $X$ and of $Y$, $n$, varies from one Mazurka to the next, and is given in Table 1.

For each Mazurka, we compute the mean Pearson correlation coefficients over all recordings of that Mazurka to produce the graph in Figure 1. The results of each machine-learning algorithm—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbour (k-NN)—is denoted on the graph using a different symbol.

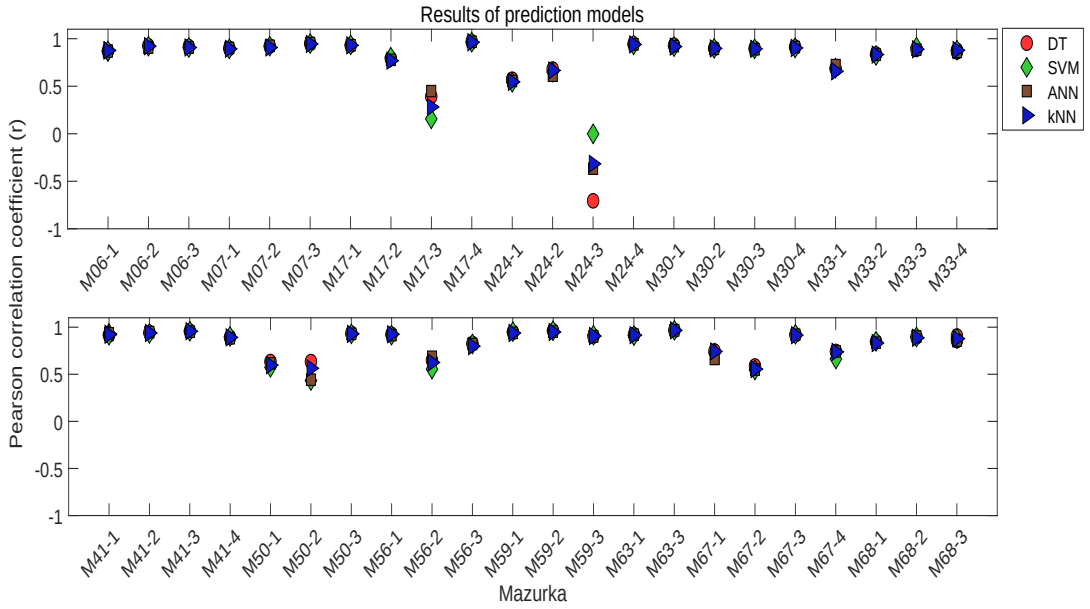


Figure 1. Pearson correlation coefficient between predicted and actual loudness values for each Mazurka, averaged over all recordings of the Mazurka, for each machine-learning method—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbour (k-NN).

Observe in Figure 1 that, with few exception, the mean Pearson correlation coefficient is fairly high. When disregarding the two most obvious outliers, Mazurka Op. 17 No. 3 and Mazurka Op. 24 No. 3, the mean Pearson correlation value ranged from 0.5192 to 0.9667 over all machine learning methods. Furthermore, the mean Pearson correlation coefficient, averaged over the four machine-learning techniques, and over all Mazurkas is equal to 0.8083. This demonstrates that, for most Mazurkas, the machine-learning methods, when trained on data from other pianists' recordings of the Mazurka, can reasonably predict the loudness choices of a pianist for that Mazurka.

In the next sections, we inspect the anomalous situation of Mazurka Op. 17 No. 3 and Mazurka Op. 24 No. 3, and the special cases when one particular pianist deviates from the behavior of other pianists for specific Mazurkas.

### *3.1.1. Cases of low correlation between predicted and actual loudness values*

While the overall Pearson correlation measure of success is good for the prediction of loudness values in a Mazurka recording when a machine-learning model is trained on other pianists' recordings of the Mazurka, two Mazurkas were found to be outliers to this positive result: Mazurka Op. 24 No. 3 and Mazurka Op. 17 No. 3.

For Mazurka Op. 24 No. 3, the mean (over all recordings) Pearson correlation value, when averaged over the four machine-learning techniques, is -0.347, meaning that the predictions are weakly negatively correlated from the actual loudness values. The mean

Pearson correlation value for Mazurka Op. 17 No. 3, when averaged over the four machine-learning methods, while positive, is low at 0.320, meaning that the predictions are only weakly correlated with the actual loudness values.

Looking deeper into the case of these two Mazurkas, apart from the common key of A♭ major, they also share the property of having only the dynamic markings ***p*** and ***mf*** in the score, with extended mono-symbol sequences of ***p***'s. In the case of Mazurka Op. 24 No. 3, the existing score markings are {***mf***, ***p***, ***p***, ***p***, ***p***, ***p***, ***p***}; for Mazurka Op. 17 No. 3, the score markings are {***mf***, ***p***, ***mf***, ***p***, ***p***, ***p***, ***p***, ***mf***, ***p***}. The narrow dynamic range of the notated symbols and the consecutive strings of the same symbols both will almost certainly lead to a wide range of interpretations in order to create dyanmic contrast and narrative interest.

Consider the case of Mazurka Op. 24 No. 3: Figure 2 shows the actual loudness values for the eight recordings at the points of the dyanmic markings. Note that, in this Mazurka, apart from the initial ***mf***, the remaining dynamic markings are uniformly ***p***. The x-axis marks the sequence of dynamic markings in the score, and the eight distinct symbols on the graphs mark the loudness values (in sones) at these points in the recording. Note the wide range of interpretation of the loudness level for ***mf***; the loudness value ranges for the ***p***'s are often as wide as that for the ***mf***, with the recordings exhibiting many contradictory directions of change from one dynamic marking to the next. In particular, note that in three out of the eight recordings, the ***p*** immediately following the ***mf*** is actually louder than the ***mf***.

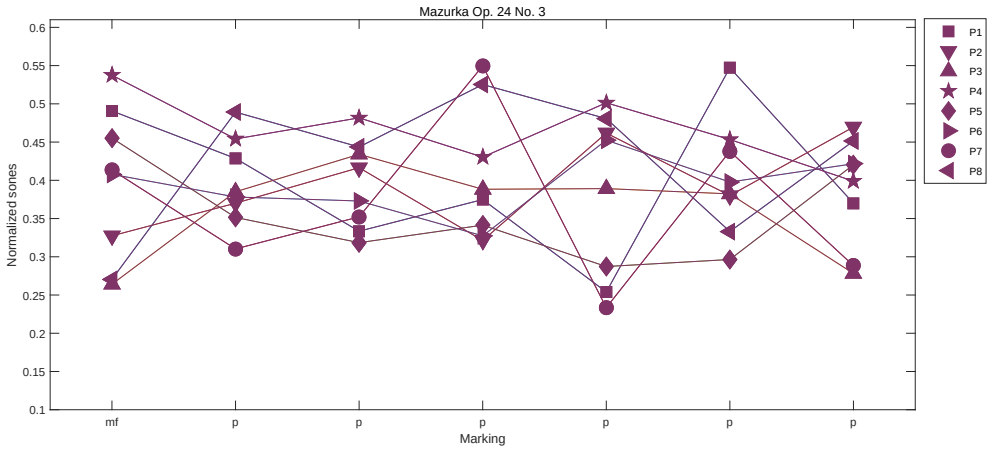


Figure 2. Representation of the loudness levels on the marking positions in score time for Mazurka M24-3 for the eight pianists.

For the case of Mazurka Op. 24 No. 3, the Pearson correlation coefficient between the predicted and actual loudness values, for each of the four machine-learning methods, are uniformly negative. Next, we consider the cases when the predictions are negatively correlated for only one recording of a Mazurka while the remaining seven recordings of the same Mazurka had predictions positively correlated with the actual loudness values.

### *3.1.2. Cases when one recording is negatively correlated while others are not*

The tests in the previous section showed that there can be a high degree of divergence in loudness interpretation amongst recordings of a Mazurka. In this section, we examine the special cases of solitary deviant behavior, when one pianist chose starkly different loudness strategies than the others when recording a Mazurka. For this, we consider

the cases when the predictions for one recording has a negative average (over all four machine-learning techniques) Pearson correlation value while those for the other seven have positive average correlation coefficients between predicted and actual loudness values.

We identified four Mazurkas for which the predictions for one recording was negatively correlated on average (over all the machine-learning algorithms) and the other seven were positively correlated. The average Pearson correlation values for each pianist's recordings of these four Mazurkas are plotted in Figure 3. The average correlation value of the solitary deviant recording for each Mazurka is highlighted with a red dot. For each red dot marking the correlation value of the worst-predicted pianist, the correlation values of the other pianists who recorded the Mazurka are shown as grey squares in the same vertical column.

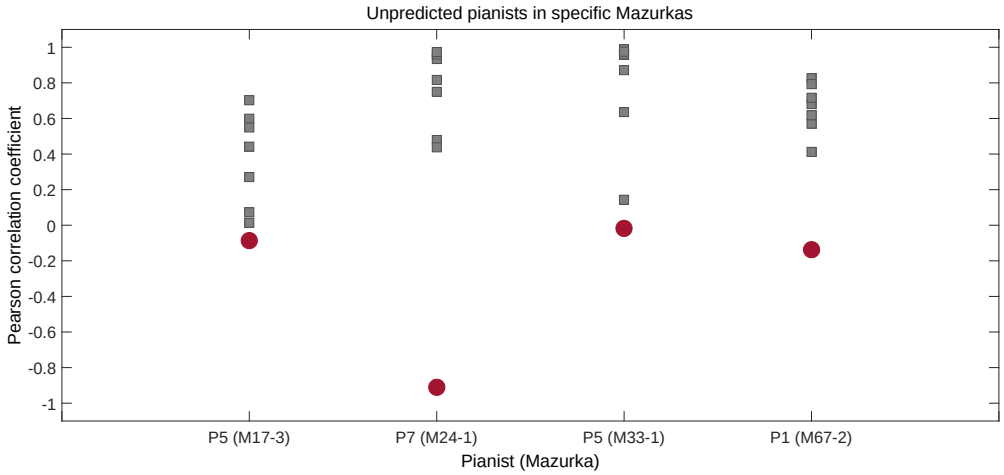


Figure 3. Pearson correlation coefficients of Mazurkas for which the worst predicted pianist's recording scored, averaging over all machine-learning methods, a negative r value (red dots): pianist P5 for M17-3, P7 for M24-1, P5 for M33-1, and P1 for M67-2. The average coefficients for the remaining pianists are shown as grey squares.

We can see from the display that even when the machine-learning algorithms did poorly in the case of a particular pianist, they often did fairly well for other pianists who recorded the same Mazurka. Figure 3 demonstrates why the loudness values of certain Mazurka recordings cannot be predicted well when machine-learning algorithms are trained on other pianists' recordings of the same Mazurka.

We next turn our attention to the extreme case of Mazurka Op. 24 No. 1, in which the loudness value predictions for one recording, that of Barbosa (P7), are amost perfectly negatively correlated with the actual values. Figure 4 shows the loudness time series, in score time, for all eight recordings of Mazurka Op. 24 No. 1. The loudness time series for Barbosa, the worst-predicted pianist, is highlighted in bold. The score dynamic markings for this Mazurka are labeled on the x-axis. The actual loudness values recorded for pianist P7 at these points are marked by red dots. The actual loudness values of other pianists——Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Czerny (P8)—at this point are marked by black symbols on the same vertical line.

As illustrated by the graph, in Barbosa's recording, he employs a performance strategy contrary to that of all or almost all of the other pianists. Furthermore, the loudness level of Barbosa's recording sampled at the points of the first four dynamic markings are relatively level, in contrast to the strategies exhibited by the other pianists.

Having considered the prediction of loudness values in a Mazurka recording by training machine-learning algorithms on recordings of the same Mazurka by other pianists, we

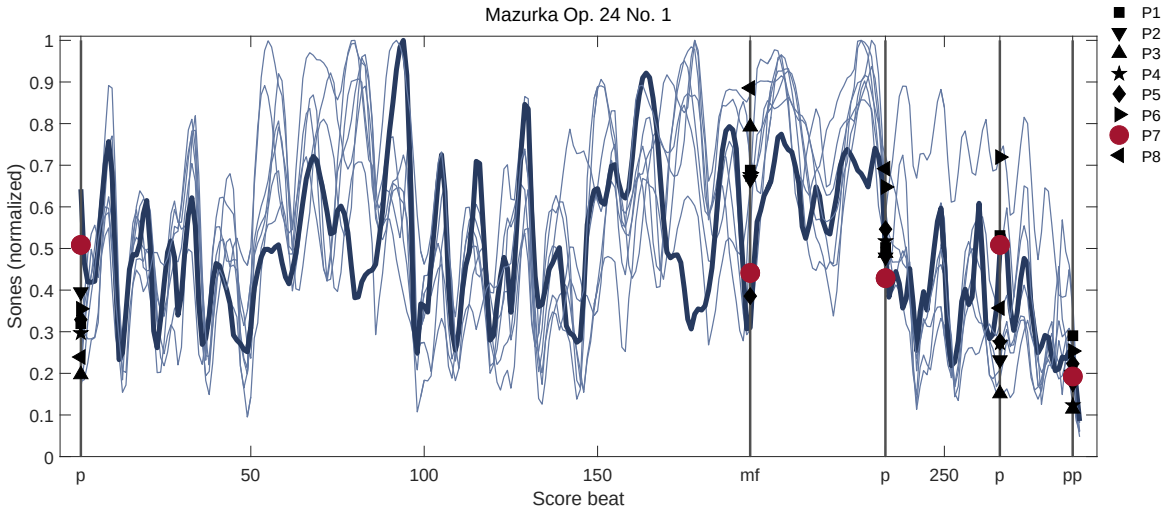


Figure 4. Loudness time series in score-beat time for recordings of Mazurka Op. 24 No. 1. Loudness time series for Barbosa's recording (P7) is shown in bold; loudness value for Barbosa's recording at dynamic markings are indicated by red dots, those of other pianists——Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Czerny (P8)—are shown as black symbols.

next consider predictions of models trained on recordings of other Mazurkas by the same pianist.

### 3.2. *Predicting loudness given target pianist's recordings of other pieces*

In the second experiment, we use the machine-learning methods to predict the loudness values at the dynamic markings of one Mazurka, given the loudness levels at the markings of other Mazurkas recorded by the same pianist. The evaluations focus on how well a pianist's loudness choices can be predicted given those of them made in the other forty-three Mazurkas. For this purpose, a 44-fold training-testing validation has been implemented. As before, we use as measure of prediction accuracy the Pearson correlation coefficient between actual and predicted values.

The results are displayed in Figure 5, which shows the mean, minimum, maximum, and median Pearson correlation coefficient values over all Mazurka recordings by each of the pianists listed in Table 2; the results are broken out into the four machine-learning methods employed—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbour (k-NN).

Contrary to the previous experiment, where the machine-learning models were trained on other pianists' recordings of the same Mazurka and did fairly well, when training the models on the same pianist's other Mazurka recordings, the average cross-validation results for each pianist are close to zero. The minimum is close to -1, implying that sometimes the loudness values of a recording can be directly contrary to the predictions, and the maximum is close to 1, implying that sometimes the loudness values behave as predicted. The results thus demonstrate that it can be extremely difficult to predict loudness values in a recording given the pianist's past behavior in recordings of other pieces. The results fare far better when training on other pianists' recordings of the same piece.

In Section 3.3, we seek to gain some insights into why the Pearson correlation values were so highly variable between the predicted and actual values for this experiment. In particular, we examine in detail the Mazurkas for which predicted and actual loudness

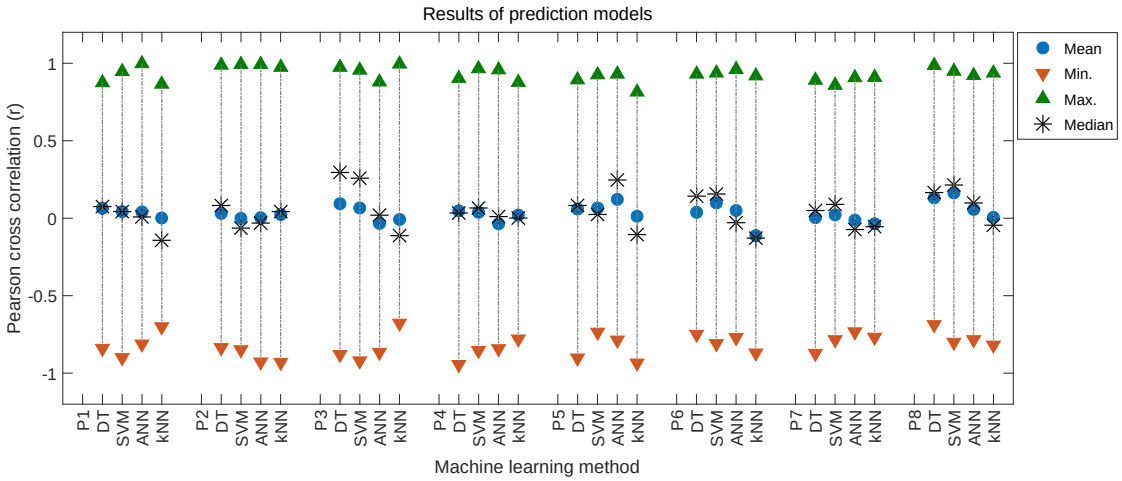


Figure 5. Pearson correlation coefficient mean, min, max, and median values for each method—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbour (k-NN)—for each pianist—Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Barbosa (P7), Czerny (P8).

values were most strongly and positively correlated and most strongly and negative correlated across all machine-learning methods.

The previous results have shown that while there may be some small variability in the prediction quality of the four machine-learning methods, they agree on the prediction difficulty amongst the recordings In the next section, we perform a second check on the prediction quality using a Euclidean measure.

### *3.2.1. Comparing machine-learning methods*

To check for variability in the prediction quality of the four machine-learning algorithms, we compute the accuracy for each algorithm using the Euclidean distance. The Euclidean distance between the predicted ($X$) and actual ($Y$) loudness values (in sones) in the held out data is given by the formula

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{10}$$

where $x_i \in X$, $y_i \in Y$, and $n$ is the size of $X$, and $Y$, which varies from one Mazurka to the next, as shown in Table 1.

We average these Euclidean distances over all Mazurkas for a given pianist to produce the results shown in Figure 6. The results are grouped by pianist: Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Barbosa (P7), Czerny (P8). For each pianist, the graph shows accuracy results for each of the four machine-learning methods—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbour (k-NN)—averaged over all Mazurka recordings by that pianist.

The results show that the algorithms perform consistently one relative to another. The span of average Euclidean distance between predicted and actual values is relatively small. In comparison, the DT algorithm produced the best results, followed closely by the SVM algorithm then the k/-NN algorithm; the ANN algorithm is a more distant

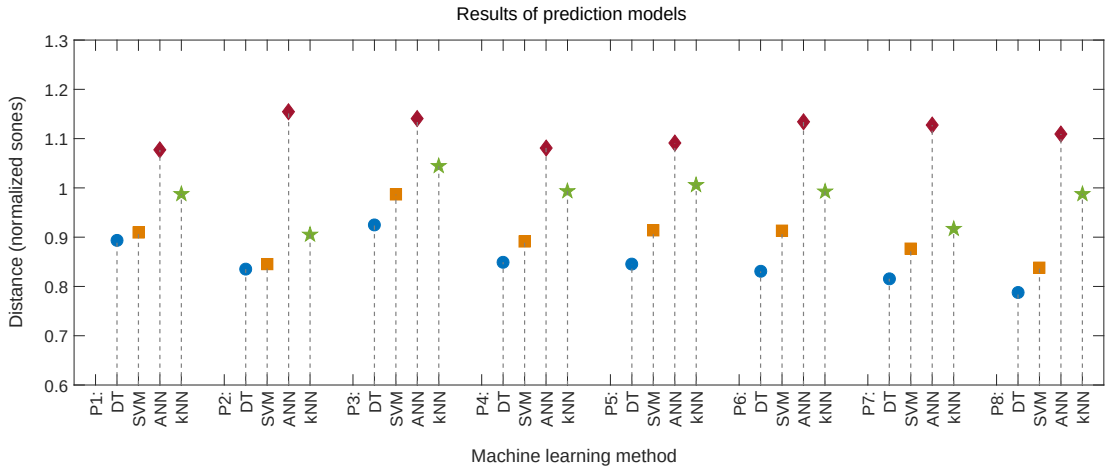


Figure 6. Euclidean distance between predicted and actual loudness values for each pianist——Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Barbosa (P7), Czerny (P8)—for each method—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbour (k-NN)—averaged over all Mazurka recordings by the pianist.

fourth. While the ANN algorithm fared worse in this exercise, we shall see in Section 4.2 that the ANN gives better results when the machine-learning algorithms are applied to the problem of predicting dynamic marking labels.

### 3.3. *A pianist's interpretation may not be predictable based on their approach to other pieces*

In this section, we dig deeper into the Pearson correlation results of the previous section to consider the extreme cases of when the predicted and actual loudness values are most consistently positively and consistently negatively correlated. We consider the Mazurkas for which all methods produced loudness predictions that were negatively correlated with the actual loudness values for all pianists; we also single out the Mazurkas for which all methods produced predictions that were positively correlated with the actual values for all pianists.

#### *3.3.1. Most negatively and most positively correlated results*

In the cross-validation, the highest Pearson correlation coefficient between predicted and actual loudness, 0.9981, is encountered in Chiu (P1)'s recording of Mazurka Op. 63 No. 3 (M63-3), and the lowest correlation value, −0.9444, in Fliere (P4)'s recording of Mazurka Op. 17 No. 2 (M17-2).

The four Mazurkas for which the Pearson correlation is negative for all the machine-learning methods and for all eight pianists are Mazurkas Op. 7 No. 1 (M07-1), Op. 24 No. 2 (M24-2), Op. 24 No. 4 (M24-4), and Op. 50 No. 3 (M50-3). This means that, for these Mazurkas, the pianists' recorded loudness strategies are contrary to those gleaned from their recordings of the other Mazurkas. The three Mazurkas for which the Pearson correlation coefficient over all pianists and for all machine-learning methods was positive are Mazurkas Op. 30 No. 4 (M30-4), Op. 41 No. 2 (M41-2), and Op. 68 No. 2 (M68-2). For these Mazurkas, the pianists' recorded loudness strategies are in accordance to those gleaned from their recordings of the other Mazurkas.

The results are summarized in Figure 7, which presents the Pearson correlation result

for each of the eight pianists for the Mazurkas mentioned; each pianist's data point for a Mazurka shows the average over the four machine-learning methods. Note that, for each Mazurka, the correlation coefficients are relatively closely grouped for all pianists. In the next section, we shall examine more closely the four Mazurkas having all-negative correlation values, i.e. the ones to the left of the dividing line.

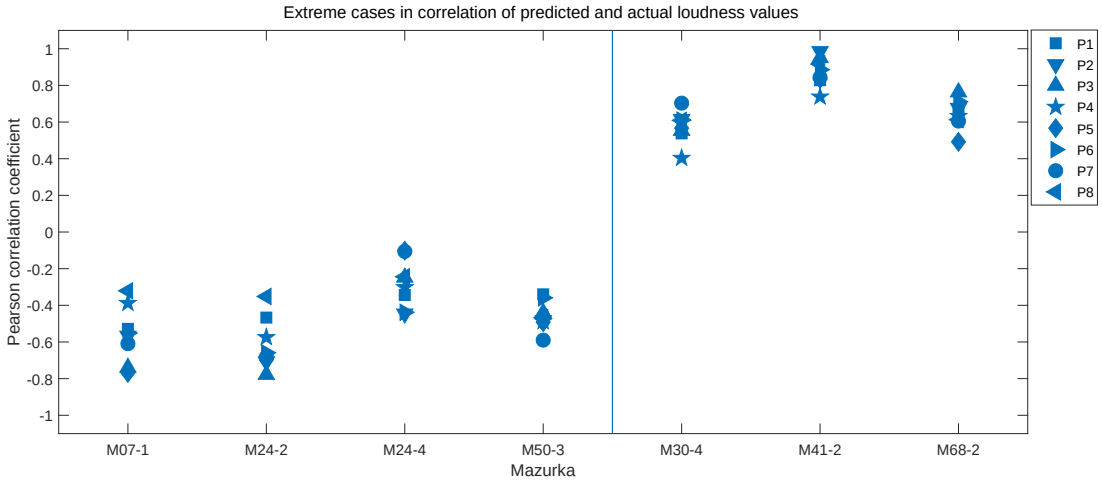


Figure 7. Pearson correlation coefficient values for each pianist, averaged over all machine-learning methods, for the Mazurkas having all negative (left: M07-1, M24-2, M24-4, M50-3) and all positive (right: M30-4, M41-2, M68-2) correlation values.

### *3.3.2. Cases of negative correlation*

Here, we focus on the four Mazurkas with negative Pearson correlation coefficients for all pianists and all machine-learning methods. Figure 8 shows the predicted and actual loudness values at each dynamic marking in the four Mazurkas, concatenated in sequence. The values shown are the average over all pianists and all machine-learning methods for each marking and Mazurka.

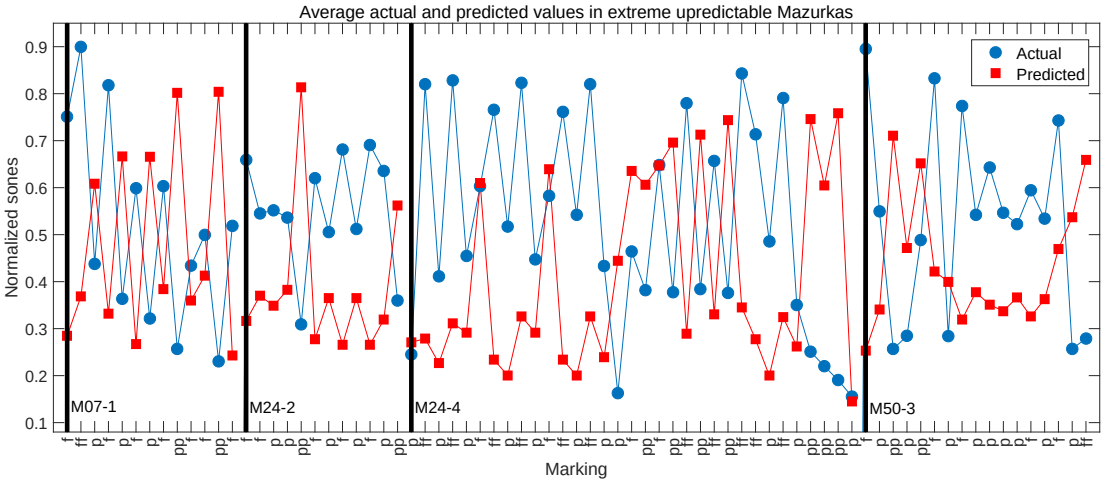


Figure 8. Average actual (circles) and predicted (squares) dynamic values at dynamic marking positions for *Mazurkas* M07-1, M24-2, M24-4, and M50-3 where the average of the Pearson correlation coefficient over pianists and methods was negative.

As can be seen, for most dynamic markings in these four Mazurkas, the actual loud-

ness levels are the opposite of the predicted loudness values. The difference in loudness strategy may be due to deliberate contrarian behavior, or to aspects of the piece relevant to dynamic choices not being captured in the selected features. A discussion on future directions for feature analysis will be presented in Section 3.5.

There may also be score-inherent factors in these four Mazurkas that contributed to the negative correlation results. These include the oscillatory nature—and hence longer range dependencies—of some of the dynamic marking sequences, and the presence of sequences of identical markings—which lead to greater variability in interpretation. Instances of oscillatory sequences include the sequence (***p***, ***f***, ***p***, ***f***, ***p***, ***f***) in Mazurka Op. 7 No. 1, the sequence (***f***, ***p***, ***f***, ***p***, ***f***, ***p***) in Mazurka Op.24 No. 2, and the sequence (***pp***, ***ff***, ***pp***, ***ff***, ***pp***, ***ff***) in Mazurka 24 No. 4. Examples of monosymbol sequences include the sequence of three ***pp***'s in Mazurka 24 No. 4, and the sequence of four ***p***'s in Mazurka 50 No. 3.

The analysis of the results of the loudness value prediction experiments leads us to suspect that the poor predictability of some of the Mazurka recordings may be due to high variability in performed loudness strategies among the pianists for the range of Mazurkas represented. Hence, in the next section, we describe and report on an experiment to test the degree of similarity in the loudness strategies employed by one pianist vs. that used by another.

### 3.4. *Inter pianist similarity*

To determine the degree of similarity in the loudness strategies employed by different pianists, machine-learning models were trained on all Mazurka recordings by one pianist and used to predict the loudness values in all Mazurka recordings by another pianist. The goodness of fit is measured using the mean (over all Mazurka recordings) of the average (over all machine-learning methods) Pearson correlation coefficient.

The results are shown in Fig. 9 in the form of a matrix, where the $(i,j)$-th element in the matrix represents the percentage of the mean averaged Pearson correlation coefficient between the loudness value predictions of the model trained on data of pianist $i$ and the corresponding actual loudness data of pianist $j$. The correlation values range from 68.36% to 78.43%. This shows the high level of similarity between pianists for interpreting dynamic markings in the pieces investigated.

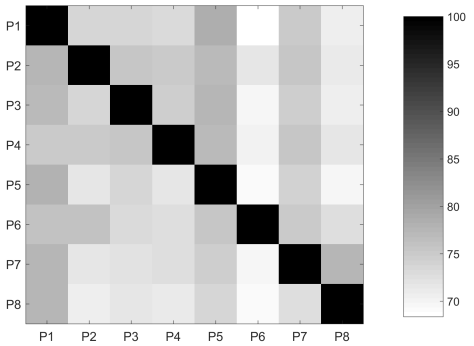


Figure 9. Matrix displaying Pearson correlation coefficient values (in %), averaged over all machine-learning methods, when applying a model trained on recordings by one pianist to predict the loudness values of another pianist.

Higher correlation values are observed in columns P1 and P5. We can deduce that Chiu (P1)'s interpretation model, followed closely by Shebanova P5's model, best predicts the loudness values of recordings by other pianists. (Note that Chiu's recordings also

achieved some of the highest correlation values, meaning that they were most predictable. Furthermore, the loudness strategies of pianists P1 and P5 best fit each other, i.e. the highest non-diagonal Pearson correlation value is found when P1's model is used to predict the loudness levels in P5's recording, and vice versa. On the other hand, the loudness strategies of pianist P6 is the one most dissimilar to that in other recordings, as shown by the almost white non-diagonal squares in column P6.

### 3.5. *Discussion on feature analysis*

The k-NN method is known to be highly sensitive to irrelevant features, i.e. it performs considerably less well than other algorithms in the presence of irrelevant features. As the results show no demonstrable trend in this respect, this leads us to think that all the extracted features in our feature set are indeed relevant.

A natural question that follows is: which of the extracted features are more salient for predicting performed loudness? To investigate this question, we apply the RELIEF algorithm from Robnik-Sikonja and Kononenko (1997) to rank the features according to relevance for predicting loudness. The RELIEF algorithm evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.

The results are shown in Table 3, with features being ranked from most important (1) to least important (8). The feature-relevance analysis indicates that the current dynamic marking proved to be the most informative feature for the task, followed by the text qualifiers, and the preceding dynamic marking. Interestingly, the ranking of features according to relevance for predicting loudness is the same for all pianists, which may indicate that all the considered pianists give the same relative importance to the investigated features when deciding loudness levels in their performances.

| Ranking | Feature | Ranking | Feature | Ranking | Feature |
|---|---|---|---|---|---|
| 1 | M | 4 | N_M | 7 | Dist_N |
| 2 | Annot_M | 5 | Dist_PR | 8 | Annot_N |
| 3 | Annot_PR | 6 | PR_M | | |

Table 3. Ranking of importance of the features for the loudness prediction task.

In order to evaluate the incremental contribution of each of the features studied, we created machine learning models using different subsets of features. More concretely, we considered feature subsets by incrementally adding features to the training set one at a time, starting with the most important feature, i.e. the highest ranked, and continuing to add features according to their rank order. In Figure 10 we present the results for each pianists recording, averaged over all machine learning methods.

As can be seen in the figure, for all pianists, the loudness prediction accuracy for their recordings, with few exceptions, increases monotonically with the number of features. This confirms our belief that all the features studied are indeed relevant and contribute to the accuracy of the loudness level predictions for all recordings. It is worth noticing that the highest ranked feature, i.e. the dynamic marking, contains on its own substantial predictive power: the correlation coefficient of the loudness models induced with only dynamic marking information alone ranges from 0.64 (for pianist P6s recordings) to 0.75 (for pianist P1s recordings).

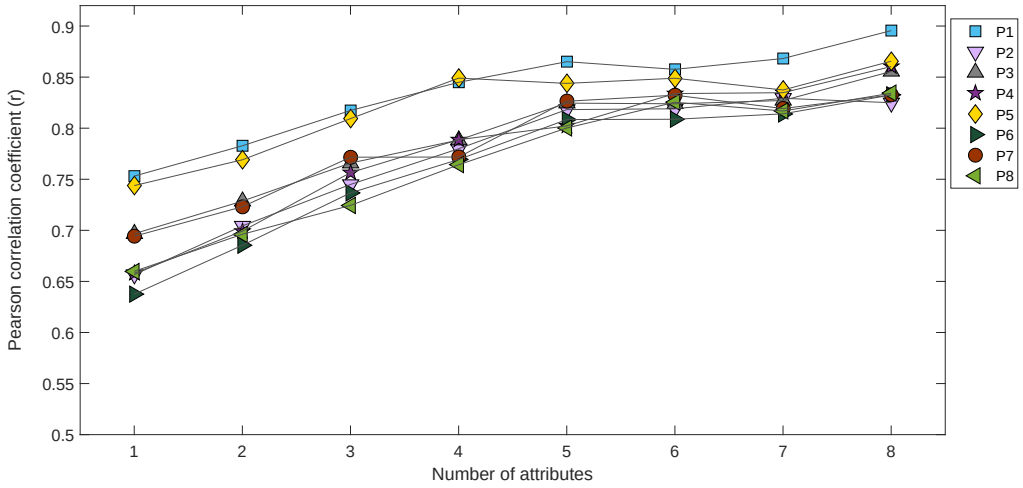


Figure 10. Upper bound for average (over all machine learning methods) Pearson correlation coefficient as number of features considered increases; features are added in the order given in Table 3. Numbers result from the optimal case where the testing data is equal to the training data.

## 4. Dynamic marking prediction

Classification remains one of the staple tasks of machine-learning algorithms. This section assesses the fit of the machine-learned models as they are applied to the classification of a Mazurka recording's loudness values into loudness categories as indicated by dynamic markings. In particular, we examine the following problem: given a loudness level in a recorded Mazurka, what is the dynamic marking that best represents conceptually the loudness value at a particular instance in time.

As in Section 3, we have conducted two experiments. In the first experiment each classification model has been trained for each Mazurka separately and a 8-fold cross-validation implemented for predicting the dynamic markings given other pianists' recordings of the target piece. In the second experiment each classification model has been trained for each pianist separately and a 44-fold cross-validation has been implemented for predicting the dynamic markings of one Mazurka given the other Mazurka recordings of the same pianist.

For evaluation, we compute the percentage of correctly classified instances. For each class of markings, {***pp***, ***p***, ***mf***, ***f*** ***ff***}, we compute the F-measure which is defined as

$$\mathrm{F} = \frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}}, \tag{11}$$

where precision, P, is the proportion of true positives for the relevant class, and recall, R, proportion of the relevant class retrieved. As the number of markings is different for each class, we have created a composite F-measure that is the sum of the F-measures for each class of markings, weighted by the proportion of markings in that class.

Therefore in the remaining part of this section we present the analysis of the results of the second experiment. More specifically, Section 4.1 presents and discusses the results for predicting dynamic markings given other pianists' recordings of the target piece; Section 4.2 does the same for predicting markings given the target pianist's recordings of other pieces; Section 4.3 looks at which of the dynamic markings are more easily classified than others; and, Section 4.4 evaluates the ease or difficulty of predicting the markings of a Mazurka using the ratio of correctly classified markings.

### 4.1. *Predicting dynamic markings given other pianists' recordings of target piece*

In the first experiment, we use the four machine-learning methods to predict the dynamic-marking labels of loudness values given the loudness-marking data of other pianists' recordings of the target piece. The cross-validation tests result in 100% correctly classified instances—and correspondingly a precision of 1.0 and recall of 1.0 for each dynamic marking class, and weighted F-measure of 1.0—for every Mazurka. This means that it is entirely plausible to train a machine-learning algorithm on other pianists' performance of a Mazurka to predict the dynamic marking labels of the target recording.

This finding, while suspicious, may not be surprising considering that the number of markings of each Mazurka that we want to predict is small, and the number of pianists is also small, fact that affects the classification results more than the results for predicting the loudness level.

We next move on to the complementary experiment in which the machine-learning algorithms predict dynamic-marking labels based on loudness-marking data of other Mazurka recordings by the target pianist.

### 4.2. *Predicting dynamic markings given target pianist's recordings of other pieces*

In this experiment, we train classifiers—using DT, ANN, SVM, and k-NN algorithms—on the loudness-label mappings of other Mazurka recordings by the same pianist in order to identify the dynamic marking corresponding to a given loudness value in the target recording. Table 4 shows the results obtained by the classifiers in terms of the mean percentage of Correctly Classified Instances (CCI) over all Mazurkas, and the mean class-weighted average F-measure (F) over all Mazurkas. The highest mean CCI values for each pianist are highlighted in bold.

| Method | | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| DT | CCI(%) | 27.822 | 26.387 | 28.012 | 24.542 | 24.798 | 27.858 | 29.230 | 26.437 | 26.8858 |
| | F | 0.264 | 0.301 | 0.309 | 0.226 | 0.210 | 0.238 | 0.320 | 0.231 | 0.2624 |
| SVM | CCI(%) | 27.735 | 25.283 | 23.962 | 22.830 | 25.283 | 24.717 | 26.227 | 25.670 | 25.2134 |
| | F | 0.262 | 0.234 | 0.230 | 0.214 | 0.246 | 0.223 | 0.229 | 0.236 | 0.2343 |
| ANN | CCI(%) | **30.755** | 27.925 | 27.736 | **26.981** | **26.604** | **31.132** | 30.189 | **30.460** | **28.9727** |
| | F | 0.323 | 0.272 | 0.258 | 0.261 | 0.280 | 0.323 | 0.303 | 0.333 | 0.2941 |
| k-NN | CCI(%) | 28.301 | **28.491** | **28.113** | 25.660 | 25.660 | 26.981 | **30.377** | 27.547 | 27.6412 |
| | F | 0.323 | 0.327 | 0.317 | 0.278 | 0.297 | 0.312 | 0.336 | 0.313 | 0.3129 |

Table 4. Percentage of Correctly Classified Instances (CCI) and weighted average F-measure per method and per pianist. Maximum values per pianist are highlighted in bold. The last column contains the average values per row and the highest average of CCI is highlighted in bold.

As can be seen by the preponderance of numbers in bold in the ANN row, the ANN algorithm gives slightly better results in terms of mean CCI than other methods. In particular, it yields the best results for pianists Chiu (P1), Fliere (P4), Shebanova (P5), Kushner (P6), and Czerny (P8), followed by the k-NN algorithm that gives slightly better results for the pianists Smith (P2), Ashkenazy (P3), and Barbosa (P7). The highest average value of the percentage of correctly classified instances over all pianists is given by the ANN algorithm. Recall that in Section 3.2.1, ANN had the lowest prediction correlation to actual figures. That was for the case of loudness prediction; here, it performed best for dynamic-marking label prediction.

The CCI's reported in Table 4 are not high, only a little above chance, which would be 20% as we consider five dynamic-marking labels. One avenue for further analysis is to identify the markings that are more easily classified by considering the ones that have been predicted correctly by all recordings; this study is reported in Section 4.3. Another direction is to identify the Mazurka which has the highest number of markings that are more easily predicted, by observing for every Mazurka the ratio of the markings correctly classified for all recordings to the total number of markings in the piece; this is reported in Section 4.4. We use the ANN algorithm as a basis for both studies due to its better performance found here.

### 4.3. *Easily predicted markings*

In this section, we seek to determine which of the dynamic-marking labels are more easily classified by the machine-learning algorithms when trained on other Mazurka recordings by the target pianist. A specific marking in a Mazurka is considered to be easily predicted if it has been classified correctly for all recordings. In Table 5 we present the markings across all Mazurkas that have been identified as being easily predicted according to this criterion.

| Mazurka | Marking (position) | Mazurka | Marking (position) | Mazurka | Marking (position) |
|---|---|---|---|---|---|
| M06-3 | ***p*** (1), ***p*** (16) | M33-2 | ***ff*** (5) | M63-1 | ***p*** (3), ***p*** (5) |
| M07-3 | ***f*** (5) | M50-2 | ***p*** (2), ***p*** (4), ***p*** (6), ***p*** (7) | M67-1 | ***p*** (4), ***p*** (15) |
| M17-3 | ***p*** (5), ***p*** (9) | M50-3 | ***p*** (12) | M67-3 | ***p*** (11) |
| M17-4 | ***ff*** (5) | M56-1 | ***mf*** (12) | M68-2 | ***pp*** (2) |
| M24-4 | ***f*** (6), ***f*** (11) | M56-3 | ***p*** (12) | M68-3 | ***p*** (2) |
| M30-3 | ***pp*** (14) | M59-3 | ***p*** (11) | | |

Table 5. Markings that have been predicted correctly for all recordings of the Mazurka containing that marking; the numbers in parentheses indicate the position of that marking in the sequence of dynamic markings in that Mazurka.

Two patterns emerge from Table 5: the first has to do with the range of the dynamic markings in the Mazurka; the second has to do with the existence of important structural boundaries such as key modulations or cadences near the marking that impact the dynamics. We shall describe each case in greater detail provide specific examples of each case.

In the first case, when considering the range of markings present in a score, the marking at the edges of this range tend to correspond to extreme loudness values in the recording. Thus, these dynamics at the extremes would be the ones most easily categorized. For example, the ***ff*** marking in Mazurka Op. 17 No. 4 (M17-4) is a case in point. The markings that appear in this Mazurka are: {***pp***, ***p***, ***ff***}. Clearly, ***ff*** is at the extreme of this marking set, and in the loudness spectrum. Even in its position in the score, it is placed uniquely in such a way as to highlight the extreme nature of its dynamic level. Figure 11 shows the score position of the marking: it is preceded by a ***crescendo*** and followed by a ***p***. It is no wonder that this marking is correctly classified in all recordings of this Mazurka.

In the second case, structural boundaries are often inflected in performance in such a way as to reinforce their existence: the dynamic may drop immediately preceding or immediately after the boundary. As an example of a dynamic marking following a key change, consider the second ***f*** marking of Mazurka Op. 24 No. 4 (M24-4), which is in the key of B♭ minor. Just before this easily predicted marking a phrase is repeated in the different key of F major, and the repeat is marked ***sotto voce*** (in an undertone). The ***f***

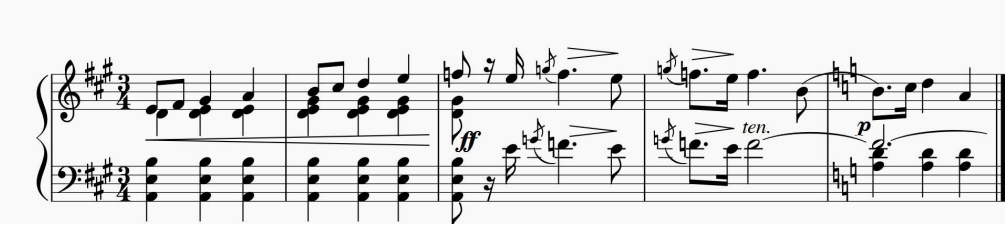


Figure 11. Case of the ***ff*** marking in Mazurka Op. 17 No. 4 (M17-4), correctly classified in all recordings of the Mazurka when the machine-learning algorithms are trained on the markings in the remaining Mazurkas by the target pianist.

appears at the return of B ♭ minor; the performer is thus encouraged to make a sharp distinction between the ***f*** and the preceding ***sotto voce***. Four bars after the ***f*** is a ***pp*** marking, which would bring into sharper relief the ***f***. These factors all conspire to make this ***f*** more easily detectable.

An example of an extreme dynamic marking near a structural boundary is the case of the marking ***pp*** in Mazurka Op. 30 No. 3 (M30-3), which is located at a cadence prior to a return to the main theme of the piece. The score segment for this example is shown in Figure 12. As can be seen in the score, the ***pp*** is proceeded by the text indicator ***dim.***, for diminuendo; furthermore, the text indicator ***slentando***, meaning to become slower, is paired with the marking; and, the marking is followed by a ***f*** marking paired with the text indicator ***risoluto***, meaning bold, at the return of the main theme. The extra meaning imputed to this ***pp*** as a signifier of the impending return of the ***f*** main theme makes it again a more extreme kind of ***pp***, and thus easier to classify.

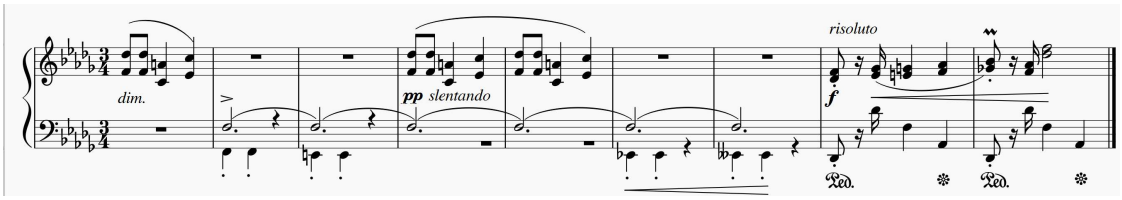


Figure 12. Case of the ***pp*** marking in Mazurka Op. 30 No. 3 (M30-3), correctly classified in all recordings of the Mazurka when the machine-learning algorithms are trained on the markings in the remaining Mazurkas by the target pianist.

In the next section, we consider the ease or difficulty of classifying dynamic markings through a different study, this time on the ratio of correctly classified markings.

### 4.4. *Easy/hard to predict Mazurkas: ratio of correctly-classified markings*

We have observed over the course of the experiments that there were *Mazurkas* for which the ratio of the markings that have been correctly classified is high for each recording, while for others that same ratio is low. To study this ratio across all *Mazurkas*, we have run a set of cross-validation experiments for which the markings of a specific Mazurka constituted the testing set and the markings of the remaining ones constituted the training set. The resulting ratio, averaged over all recordings of the Mazurka, of correctly classified markings to total markings is laid out in Figure 13 in a monotonically increasing order.

Note that Mazurka Op. 24 No. 1 has an average ratio of zero, meaning that none of its markings are classified correctly for any recording. Recall that Mazurka Op. 24 No. 1 was also the one in which the loudness value predictions for one recording was

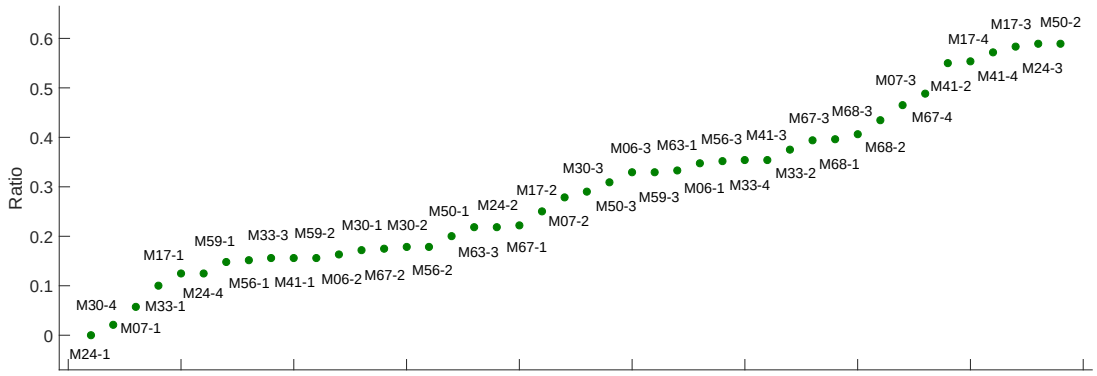


Figure 13. Average ratio, over all machine-learning methods, of correctly classified markings and number of markings per Mazurka over all pianists.

almost perfectly and negatively correlated with the actual values; this was described in Section 3.1.2.

In contrast, Mazurka Op. 50 No. 2 has the highest ratio, 0.5893, meaning that this Mazurka has the highest number of correctly classified markings for every recording. We then consider in detail the loudness values at dynamic markings in Mazurka Op. 50 No. 2 for all eight recordings of that Mazurka. In Figure 14 the markings that are correctly classified for all recordings in are highlighted with a solid vertical line, while the ones that are correctly classified for seven out of eight recordings are highlighted with a dashed vertical line. The loudness levels at these easily-classified markings follow the patterns established by other recordings.

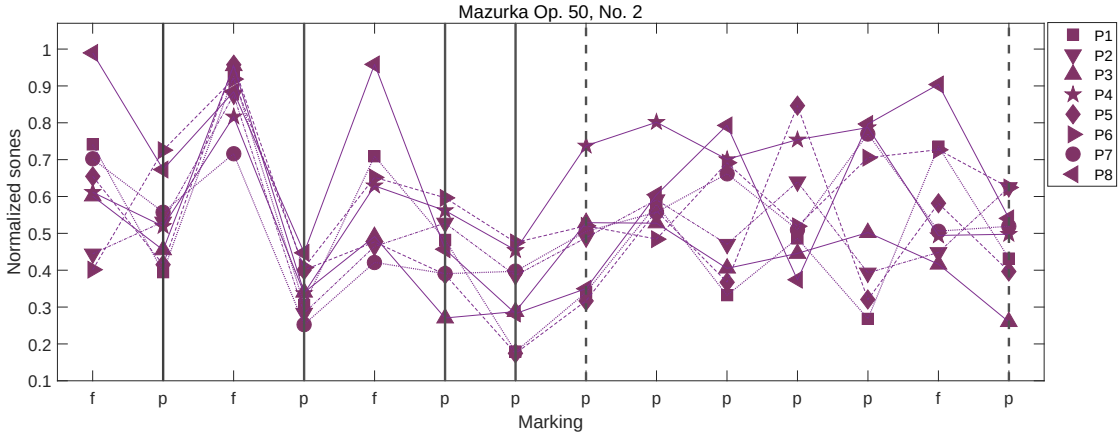


Figure 14. Loudness values at dynamic marking positions for Mazurka Op. 50 No. 2. Solid vertical lines indicate markings that are correctly classified for all eight recordings; dashed vertical lines indicate markings that are correctly classified for seven out of eight recordings.

Recall that consecutive strings of the same symbol posed significant challenges to the prediction of loudness values described in earlier sections. Note that Mazurka Op. 50 No. 2 again features a series of seven ***p***'s, with high degrees of variation in the loudness levels at these markings and in the ***f*** immediately following the sequence.

## 5. Conclusions

In this article We have investigated the relationship between loudness levels and dynamic markings in the score. To this end, we have implemented machine-learning approaches for the prediction of loudness values corresponding to different dynamic markings and musical contexts, as well as for the prediction of dynamic markings corresponding to different loudness levels and musical contexts. The methods—Decision Trees, Support Vector Machines, Artificial Neural Networks, and a k-Nearest Neighbor algorithm—are applied to forty-four recordings of performances of Chopins Mazurkas each by eight pianists.

The results in Section 3 highlight that, using any of the methods, loudness values and markings can be predicted fairly well when training across recordings of the same piece, but fail dismally when training across recordings of other pieces by the same pianist. This happens possibly because the score is a greater influence on the performance choices than the performers individual style for this test set. More specifically, insights from the data structure come forward, including Mazurkas in which loudness values can be predicted easier than others. This finding is related to the range of the different markings that appear in a piece as well as their position and function in the score in relation to structurally important elements.

The results in Section 4 highlight the good performance of the methods on predicting the dynamic markings given the loudness-marking data of the pianists' recordings of a target piece. When training across recordings of other pieces by the same pianist, the results, while not exceptional with respect to the prediction of dynamic markings, show notable trends on markings that are classified correctly, and on pieces that have higher ratios of classified markings over all markings. These trends are based mostly on the relationship between the position of the markings and the structure of the piece.

Different tuning of the parameters in the existing machine learning techniques for the prediction of loudness levels as well as for the prediction of dynamic markings may give better results. The scope of this article, however, is not to find the optimal solution to the tuning issue, but to point out universal characteristics of this kind of complex data. Improvements of the results may occur by considering possible alterations in the feature set or the data set. From the features set point of view, it would be especially interesting if the results were to improve with more score-informed features, and features that correspond to changes of other expressive parameters, such as tempo variations. From the data set point of view, a selection of pieces that include specific different sequences of markings, or a more constant range of markings could enhance the results.

Modeling expressive performance trends using artificial intelligence appears to be a rather delicate matter. The models are limited by the particular representations of changes that happen throughout a music piece with respect to the expression rendering. The results that are drawn from the current study are likely to appear in similar studies for other instruments and data sets. Future work in this direction should include the comparison of performances from different eras, or performances of musicians from particular schools. At the same time it should be kept in mind that the freedom of employing a particular dynamic range in one interpretation is a matter of the performer's idiosyncrasy.

## Acknowledgements

The authors would like to thank the anonymous referees, and especially Professor Thomas Fiore, Editor-in-Chief, for their immensely helpful suggestions and comments that substantially strengthened their original submission.

## Funding

This work was supported in part by UK EPSRC Platform Grant for Digital Music (EP/K009559/1), the Spanish TIN project TIMUL (TIN2013-48152- C2-2-R), and the European Unions Horizon 2020 research and innovation programme under grant agreement No 688269.

## References

Chauvin, Yves, and David E. Rumelhart. 1995. *Backpropagation: theory, architectures, and applications.* Psychology Press.

Cristianini, Nello, and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods.* New York, NY, USA: Cambridge University Press.

Ewert, Sebastian, Meinard Müller, and Peter Grosche. 2009. "High resolution audio synchronization using chroma onset features." In *thirty-fourth IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Taipei, Taiwan, 1869–1872.

Fabian, Dorottya, Renee Timmers, and Emery Schubert. 2014. *Expressiveness in music performance: Empirical approaches across styles and cultures.* Oxford University Press.

Grachten, Maarten, and Florian Krebs. 2014. "An Assessment of Learned Score Features for Modeling Expressive Dynamics in Music." *IEEE Transactions on Multimedia* 16 (5): 1211–1218.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. "The WEKA Data Mining Software: An Update." *SIGKDD Explor. Newsl.* 11 (1): 10–18.

Khoo, Hui Chi. 2007. *Playing with Dynamics in the music of Chopin, Ph.D. thesis.* Royal Holloway, University of London.

Kosta, Katerina, Oscar F. Bandtlow, and Elaine Chew. 2014. "Practical Implications of Dynamic Markings in the Score: Is Piano Always Piano?." In *Audio Engineering Society (AES) 53rd International Conference on Semantic Audio,* London, UK.

Paderewski, I. J., L. Bronarski, and J. Turczynski. 2011. *Fryderyk Chopin, Complete workds, X Mazurkas, twenty-ninth edition.* Fryderyka Chopina, Polskie Wydawnictwo Muzyczne SA.

Quinlan, John Ross. 1986. "Induction of Decision Trees." *Machine learning* 1: 81–106.

Robnik-Sikonja, Marko, and Igor Kononenko. 1997. "An Adaptation of Relief for Attribute Estimation in Regression." In *Proceedings of the Fourteenth International Conference on Machine Learning,* ICML '97, San Francisco, CA, USA, 296–304. Morgan Kaufmann Publishers Inc.

Sapp, Craig. 2008. "Hybrid Numeric/Rank Similarity Metrics for Musical Performance Analysis." In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR),* Philadelphia, USA, 501–506.

Widmer, Gerhard, and Werner Goebl. 2004. "Computational Models of Expressive Music Performance: The State of the Art." *Journal of New Music Research* 33 (3): 203–216.